PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data

Supplementary Material

Yanxiao Zhang[1,*], Yu-Hsuan Lin[1,*], Timothy D. Johnson[2], Laura S. Rozek[3], Maureen A. Sartor[1,2, †]

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Biostatistics and [3]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, United States

Outline of supplementary Materials:

Supplementary Materials

Versions and detailed parameters of the programs used (commands in grey):

All programs were run under default parameters if possible. Significance cut-offs for all programs are provided in Table S4. The shift size and window size estimated for each dataset with PePr are provided in Table S5. For Separate Analysis (SA) approaches, the final peak regions were defined as the intersection of the peaks generated from all separate runs; the significance of each peak was defined as the average of the ranks in all separate analyses. See details below:

PePr version 1.0.1: default parameters were used. For TFs:

PePr –c chip_file –i control _file –f file_format --peaktype=SHARP –remove_artefacts

For H3K27me3:

PePr –c chip_file –i control _file –f file_format --peaktype=BROAD

MACS version1.4.0rc2: default parameters were used.

macs14 -t chip_file -c control_file

SPP version 1.10.1: SPP was run with the ENCODE project IDR guidelines. IDR thresholds of 0.01 and 0.0025 were chosen for the original replicate threshold and pooled-pseudoreplicate threshold, respectively. The optimum set was reported.

Rscript run_spp.R -c=chipSampleRep1.tagAlign.gz -i=controlSampleRep0.tagAlign.gz -npeak=300000 -odir=/peaks/reps -savr -savp -rf -out=/stats/phantomPeakStatsReps.tab

MACS2 version 2.0.10.09132012: MACS2 was run with the ENCODE project IDR guidelines. IDR thresholds of 0.01 and 0.0025 were chosen for the original replicate threshold and pooled-pseudoreplicate threshold, respectively. The optimum set was reported.

macs2 callpeak -t chipSampleRep1.tagAlign.gz -c controlSampleRep0.tagAlign.gz -f BED -n chipSampleRep1_VS_controlSampleRep0 -g hs -p 1e-3 --to-large

ZINBA version 2.01: The alignability function was run with the corresponding genome and read mappability file. A default read extension of 90 was used. For histone data, the "broad" argument was given.

generateAlignability(…,athresh=1,extension=90),

zinba(…, refinepeaks=0,seq=chip_file, input=control_file, filetype="bed", extension=90)

SICER version 1.1: Parameters recommended by the manual were used. The corresponding genome was used for each experiment. A window size of 200 and gap size of 600 were used for broad peaks. The fragment size was set to 150.

SICER.sh chip_file control_file .genome 2 200 150 0.8 600 1E-2

edgeR version 3.2.4. *edgeR-basic*: First the reads were shifted (45bp) and counted in non-overlapping windows (200bp). The read counts and group assignments were prepared in edgeR specified format and then the following commands were applied. Windows passing the significance cut-off were deemed eligible and then adjacent windows were merged to form a final peak list. *edgeR-plus*: All of PePr's pre-processing (shift size and window size estimates) and post-processing steps (removing artefacts) were applied using our default parameter settings.

```
y = DGEList(counts=counts,group = group)
y = calcNormFactors(y)
y<-estimateCommonDisp(y, rowsum.filter=5)
y<-estimateGLMTagwiseDisp(y,design)
fit_tag<-glmFit(y,design)
lrt.tagwise<-glmLRT(fit_tag,coef=2)
```

DiffBind version 1.10.0: First, SICER was used to call peaks from each sample using the matching input samples as controls. The resulting peak lists from all four samples were input to DiffBind, which generated 29510 pre-candidate regions. The following commands were then executed to search for differential binding regions:

```
hpv = dba(sampleSheet="diffbind_sample.csv")
hpv  = dba.count(hpv)
hpv = dba.contrast(hpv,hpv$masks$`HPV-`, hpv$masks$`HPV+`, "HPV-", "HPV+")
```

For DESeq:

```
hpv =dba.analyze(hpv,method=DBA_DESEQ)
hpv.DB = dba.report(hpv,method=DBA_DESEQ)
```

And for edgeR:

```
hpv =dba.analyze(hpv, bReduceObjects=F)
hpv.DB = dba.report(hpv)
```

diffReps version 1.55.4: Default parameters. An exact negative binomial test was used. Settings were slightly different between TFs and H3K27me3.

For transcription factors, the sharp option was enabled and a window size of 200 was used:

```
diffReps.pl –tr chip_file1 chip_file2 … -co control_file1 control_file2 … -gname genome –me nb –nsd sharp –window 200
```

For H3K27me3, the default parameters were used (nsd="broad" and window size 1000):

```
diffReps.pl –tr chip_file1 chip_file2 … -co control_file1 control_file2 … -gname genome –me nb –nsd
```

Supplementary methods for PePr:

Input read file formats

PePr currently supports multiple read alignment formats, including SAM, BAM, BED, ELAND_MULTI and ELAND_EXTENDED. PePr can analyze two ChIP group comparisons or a group of ChIP samples versus control(s), e.g. input samples. The number of samples in each group can be different.

Preprocessing of data

Removal of duplicated reads (optional): For every sample, PePr offers the option to remove the duplicated reads mapped at the same genomic location. Sometimes the same DNA fragments can be sequenced repeatedly due to PCR amplification or library preparation and are over-represented in the library. Assuming each piece of DNA in the genome has equal probability of being sequenced, then the occurrence of the same sequence read multiple times would be low and would depend on the sequencing depth. Therefore, PePr removes extra duplicated reads that are beyond the expected maximum at each genomic location. The maximum is calculated using a binomial distribution as specified in (Zhang, et al., 2008).

Fragment length estimation: PePr estimates the shift size (half of the DNA fragment length) for each ChIP sample and shifts all reads to their 3' direction by this amount. For single-end ChIP-Seq data, since the sequencing read length is shorter than the DNA fragment length, the cluster of forward-strand reads and that of reverse-strand reads at the binding sites show a phase lag. Properly shifting both strands of reads towards the center of the DNA fragment can improve the power and precision of detecting binding sites. For each chromosome, PePr shifts all the reads by several attempted shift sizes (starting at zero and increasing base by base), and counts the overlap between reads from forward and reverse strands. The shift size which maximizes the overlap is the optimum shift size. For every ChIP sample, the median of the estimated shift sizes from five chromosomes (chr1 to chr5) is calculated and used. We use the median from five chromosomes to balance speed and robustness against potential outliers; the shift sizes estimated from these five chromosomes have been consistent for all datasets tested thus far. If control samples are included in the analysis, the average shift size derived from the ChIP samples are applied to the controls.

Window size estimation: To divide the genome into windows, a recommended window size is calculated as the estimated average width of the peaks, allowing PePr to optimally capture the reads in peak regions. To achieve this goal, we first divide the genome into non-overlapping 20bp bins. For each chromosome, the bin with the largest number of reads is chosen as the seed and extended to the flanking bins until a bin is reached which has less than 10% of the reads in the seed bin. The combined width of these bins is recorded. The abovementioned process is repeated 100 times after which the median of the widths is calculated. The median of widths for all chromosomes is the recommended window size. The genome is then divided into windows of the chosen (either recommended or user-specified) size that overlap by 50% and the number of reads in each window is multiplied by the normalization constant for each sample.

Normalization: The total number of reads often varies among samples, and the immunoprecipitation efficiency can also differ substantially among ChIP samples, which may artificially increase the variation among samples if unnormalized, raw read counts are used. Currently, PePr uses the Normalization of

ChIP-Seq (NCIS) method (Liang and Keles, 2012) to normalize input (control) samples and a modified Trimmed Mean of M values (TMM) method to normalize ChIP samples(Robinson and Oshlack, 2010).

First, PePr splits the genome into 1000bp bins. The mean of all ChIP libraries is used as the reference sample, towards which every sample will be normalized. For every input sample ***i*** versus the reference ***r***, let n_{ig} and n_{rg} be the number of reads in the g^{th} genomic bin for the input and reference samples, respectively. The normalization factor $\hat{r}$ for the input sample is calculated as

$$\hat{r} = \frac{\sum_{g \in B} n_{rg}}{\sum_{g \in B} n_{ig}}. \quad (1)$$

where *B* represents the background bins (in which no enrichment by the antibody exists). Let $n_g = n_{rg} + n_{ig}$, Given background bins are more likely to have lower numbers of reads, we define $B = \{g: n_g \leq \hat{t}\}$, where the count threshold $\hat{t}$ is the smallest t for which *B* consists of > 0.75 of the genome; this percentage was used and tested in (Liang and Keles, 2012), and works well as long as the DNA binding protein does not bind to >25% of the genome. Finally, the number of reads in each window for the input sample is multiplied by its normalization factor, $\hat{r}$. The process is then repeated for each input sample.

To normalize the ChIP samples for different immunoprecipitation efficiencies, for each ChIP sample, ***c***, versus the reference ***r***, the bin-wise log fold change for the g^{th} genomic bin is defined as

$$M_g = log_2(\frac{n_{rg}}{n_{cg}}) \quad (2)$$

and the geometric mean of log read counts is defined as

$$A_g = \frac{1}{2} log_2(n_{rg} \cdot n_{cg}). \quad (3)$$

Where n_{rg} and n_{cg} are the raw read counts in the g^{th} bin of the reference sample and target ChIP sample, respectively. The trimmed mean of M_g values (TMM) is calculated as the weighted average of M_g after removing the upper and lower *x* percentages of data (based on both M_g and A_g) as described in (Robinson and Oshlack, 2010). The default trimming percentages for M_g and A_g are 20% and 5% respectively. For M_g, 20% is a conservative estimate to exclude the differential sites, whereas for A_g, 5% is used to remove the highest and lowest signal regions where there may be a high percent of artefacts. The log fold change, M_g, is weighted by the mean log read counts. Thus,

$$log_2(TMM_c) = \frac{\sum_{g \in G^*} A_g M_g}{\sum_{g \in G^*} A_g} \quad (4)$$

where G^*denotes the remaining bins after the trimming. Since we aim to normalize for the difference in antibody efficiency among the ChIP samples, the normalization constant should be estimated only from enriched regions. Inclusion of background bins will bias the estimator towards the library ratio (e.g., if all bins were used the estimator would equal the library total read count ratio). In practice, the number of enriched regions varies across different TFs and it may not be clear how many bins should be included before we have formally called the peaks. To overcome this uncertainty, PePr sorts the bins by $n_g = n_{rg} + n_{cg}$ and estimates the TMM from the largest N bins, where N is a vector of values ranging from 1,000 to 50,000 (1000, 5000, 10000, 20000, 30000, 40000, 50000; the range was set based on the number

of peaks observed for common TFs and histones). From the several TMMs estimated from the different Ns, the one that is most different from the library ratio is reported. This will be close to the optimal TMM because as N increases toward the true number of peaks, the TMMs trend away from the library ratio, approach the enrichment signal ratio, and then eventually return to converge to the library ratio as N surpasses and grows beyond the true number of peaks. Plots illustrating the steps of this normalization process are available on our website at http://code.google.com/p/pepr-chip-seq/.

Detection of significant windows

Read counts in the test and control sample groups (or two ChIP sample groups) are modeled using the negative binomial distribution as described here.

Let Y_{ijk} denote the observed number of reads in the i^{th} genomic window(i=1,…,I), the j^{th} replicate(j=1,…,J_k) and k^{th} group(k=1,2). Assuming a negative binomial distribution, we have

$$Y_{ijk} \sim NB\left(\mu_{i\cdot k}, \varphi\right) = \frac{\Gamma\left(y_{ijk}+\varphi^{-1}\right)}{\Gamma(\varphi^{-1})\Gamma\left(y_{ijk}+1\right)} \frac{{\varphi^{-1}}^{\varphi^{-1}} {\mu_{i\cdot k}}^{y_{ijk}}}{\left(\varphi^{-1}+\mu_{i\cdot k}\right)^{\left(\varphi^{-1}+y_{ijk}\right)}} \quad (5)$$

where $\mu_{i\cdot k} = E\left(Y_{ijk}\right)$($j$=1,..,$J_k$)and φ is the dispersion factor (as $\varphi \rightarrow 0$, the distribution converges to a Poisson distribution). By parameterizing the means of read counts for each window i as $\mu_{i\cdot 1} = \mu_{i\cdot}$and $\mu_{i\cdot 2} = \gamma\mu_{i\cdot}$, we can test for a significant difference between two groups by testing the following hypothesis:

$$H_0: \gamma \leq 1 \; vs \; H_1: \gamma > 1$$

In the case of test (ChIP) versus control comparisons, the controls are assigned as group 1 and test samples are group 2 so only one direction of the hypothesis will be tested; whereas in the case of two ChIP group comparisons (i.e. differential binding), a sample/group swap is performed and the hypothesis is tested both ways automatically.

The local dispersion parameter is estimated for each window using a weighted average of initial dispersion estimates from local windows in order to gain more robust estimates as described here. The log-likelihood for a given window is:

$$l_i(\varphi) = \sum_{k=1}^{2}\sum_{j=1}^{J_k}\left[\log\Gamma\left(y_{ijk}+\varphi^{-1}\right) - log\Gamma(\varphi^{-1}) - log\Gamma\left(y_{ijk}+1\right) + \varphi^{-1}\log(\varphi^{-1}) + y_{ijk}\log\left(\widehat{\mu_{i\cdot k}}\right) - \left(\varphi^{-1}+y_{ijk}\right)\log(\varphi^{-1}+\widehat{\mu_{i\cdot k}})\right] \quad (6)$$

$$\text{where } \widehat{\mu_{i\cdot k}} = \frac{\sum_{j=1}^{J_k} y_{ijk}}{J_k}$$

The local dispersion estimator $\widehat{\varphi}$ maximizes the log likelihood over W nearby windows (including the current window) using the triangular weight:

$$L(\varphi) = \sum_{x=-W}^{W} \left(1 - \frac{|x|}{w+1}\right) l_{i+x}(\varphi) \quad (7)$$

The use of a local dispersion estimator provides a stable estimator of the dispersion factor when the sample size is small. *W* is one for the SHARP peak setting and ten for the BROAD peak setting, based on observations of autocorrelation in multiple datasets. To calculate the significance, we use an asymptotic Wald's test with log transformation. We can define:

$$Z_i = \frac{[g(\hat{\gamma}) - g(\gamma_0)]}{g'(\hat{\gamma})\widehat{\sigma_{\hat{\gamma}}}} = \frac{[\log(\hat{\gamma}) - \log(\gamma_0)]\hat{\gamma}}{\widehat{\sigma_{\hat{\gamma}}}} \quad (8)$$

Where Z_i has an asymptotic standard normal distribution, $\hat{\gamma} = \bar{y}/\bar{x}$, $\gamma_0 = 1$, and $\widehat{\sigma_{\hat{\gamma}}}$is defined as

$$\widehat{\sigma_{\hat{\gamma}}} = \sqrt{\frac{\bar{y}[J_1\bar{x}(\hat{\varphi}^{-1} + \bar{y}) + J_2\bar{y}(\hat{\varphi}^{-1} + \bar{x})]}{J_1 J_2 \hat{\varphi}^{-1}\bar{x}^3}} \quad (9)$$

where$\bar{x} = \frac{\sum_{j=1}^{J_1} y_{ijk}}{J_1}$ and $\bar{y} = \frac{\sum_{j=1}^{J_2} y_{ijk}}{J_2}$

P-values are calculated using Z_i as the test statistic. Windows satisfying the specified p-value cutoff (the default is1e-5) are called as significant windows. Benjamini-Hochberg FDR is also reported.

Defining peak regions and post-processing of peaks.

The significant windows that are localized in the same genomic area are merged. PePr has two different settings for merging windows; the maximal merging distance is smaller for the SHARP peak setting and larger for the BROAD peak setting (to ensure that the broad histone peaks are not broken into multiple enrichment regions in a given area). Generally in an explorative analysis when the enrichment shape of the peak is unknown to the user, the latter BROAD peak setting is recommended.

Optionally, PePr can remove peaks due to a high level of PCR duplicates in ChIP samples. Those peaks show no strand lag between forward and reverse strand reads (Landt, et al., 2012) and are very likely to be false positives; a high proportion of these peaks in the final peak list is an indicator of poor data quality. Removing these artifacts requires accurate estimation of the shift size, otherwise we will be risking removing true positives. Fortunately, these artifacts also occur in a properly prepared control sample, displaying similar read profiles. Thus, PePr tackles this issue by removing peaks that have similar shape in both the ChIP and input samples. Specifically, for each peak, let π_{xk} be the proportion of reads in the peak at nucleotide position *x* for group *k*, where *k*=1 is the ChIP group and *k*=2 is the input group. Reverse-strand reads are counted at their 3' end. Thus, for each group *k*, $\sum_{x \in P} \pi_{xk} = 1$, where *P* is the entire set of positions in the peak region.. The minimum of the ChIP and input proportion at each position is determined, and the resulting values are summed to define the value R across all positions in the peak using the formula:

$$R = \sum_{x \in P} \min(\pi_{x1}, \pi_{x2}) \quad (10)$$

R ranges from 0 to 1, and will have a high value when the peak shape is similar between ChIPs and controls; based on observations, technical artifacts typically have a high R value greater than 0.5, whereas R values for most (true) peaks are distributed between 0 and 0.2. PePr removes the peaks having R value greater than 0.5.

Additionally, PePr evaluates the overlap between forward-strand reads and reverse-strand reads (counted at their 3' end) before and after shifting. A peak with strand-overlap-ratio that is high (> 0.2) before shifting and decreases significantly after shifting (decrease >50% of the original level) is removed by PePr. Most PCR-duplicate peaks simultaneously meet both of the two criteria defined above. These removed peaks are reported in a separate file.

Finally, PePr offers the option to refine the peak width for sharp peaks. Typically for TFs, downstream analysis such as motif analysis works optimally with a fine resolution of the peaks (i.e. reduced to minimal width that may contain the core protein-protected binding region). In an ideal (hypothetical) ChIP experiment, the core DNA binding site would be between the last starting position of the forward-strand reads and the first starting position of the reverse-strand reads. However, real-life ChIP-Seq experiments are "contaminated" substantially by the background sequences (the percentages were observed to vary from 30% to close to 100% of the library (Liang and Keles, 2012)) and complicated by other technical factors influencing mappability and sequencability. Therefore, we use a more robust method to narrow the peak width without losing the protected region by setting the left boundary to be at the 20% quantile of the starting position from the forward-strand reads and the right boundary to be at the 80% quantile of the starting position from the reverse–strand reads. These percentages are conservatively chosen.

Differential peak binding: the differential binding analysis entails an extra step compared to the peak calling analysis. In addition to the routine pre-processing steps, the reads in each window of an input sample will be subtracted from its respective paired ChIP sample if they are matched. In the case of uneven number of ChIP/input samples within each group or unpaired ChIP and input samples, the mean input reads will be subtracted from each ChIP sample. Any negative resulting values are redefined as zero counts. As mentioned earlier, the hypothesis will be tested both ways, calling differential binding sites enriched in each group.

Table S1. ENCODE data descriptions

Dataset	Cell line	Genome	Source
NRSF*	K562	hg18	HudsonAlpha
CTCF	GM12878	hg19	Broad
GABP	GM12878	hg19	HudsonAlpha
NRF1	GM12878	hg19	Stanford
SMC3	GM12878	hg19	Stanford
USF1	GM12878	hg19	HudsonAlpha
USF2	GM12878	hg19	Stanford

*Note: NRSF from GM12878 (hg19) was also tested; however, it was of poor quality, and had a large percent of peaks with zero shift size between strands).

Table S2. Total number of mapped reads in each sample.

Dataset	Sample	Number of reads (Millions)	Sample	Number of reads (Millions)
NRSF	ChIP rep1	16.1	Control rep1	16.3
	ChIP rep2	26.6	Control rep2	14.3
ATF4	ChIP rep1	28.4	Control rep1	26.8
	ChIP rep2	30.0	Control rep2	27.5
	ChIP rep3	30.2	Control rep3	28.2
CTCF	ChIP rep1	11.5	Control rep1	7.1
	ChIP rep2	20.2	Control rep2	6.4
GABP	ChIP rep1	18.4	Control rep1	5.0
	ChIP rep2	29.3	Control rep2	4.2
NRF1	ChIP rep1	12.2	Control rep1	5.9
	ChIP rep2	25.7	Control rep2	5.0
SMC3	ChIP rep1	22.8	Control rep1	5.9
	ChIP rep2	27.4	Control rep2	5.0
USF1	ChIP rep1	17.0	Control rep1	5.0
	ChIP rep2	30.1	Control rep2	4.2
USF2	ChIP rep1	15.7	Control rep1	5.9
	ChIP rep2	14.2	Control rep2	5.0
H3K27me3	HPV(+) ChIP rep1	72.2	HPV(+) control rep1	86.8
	HPV(+)ChIP rep2	85.0	HPV(+)control rep2	77.9
	HPV(-) ChIP rep1	91.3	HPV(-) control rep1	66.5
	HPV(-) ChIP rep2	79.4	HPV(-) control rep2	83.6

Table S3. Total number of peaks identified in each TF dataset.

Method \ Dataset	NRSF	ATF4	CTCF	GABP	NRF1	SMC3	USF1	USF2
PePr	5,284	15,338	34,548	5,158	4,729	25,789	6,837	5,025
MACS-CR	15,068	39,774	50,286	5,920	13,052	48,945	36,517	26,755
ZINBA-CR	9,468	25,684	57,398	5,880	14,052	62,044	12,343	23,376
MACS-SA	4,495	10,592	38,576	3,122	5,344	15,861	5,777	7,476
ZINBA-SA	5,374	11,453	41,675	4,613	6,286	21,912	5,706	9,060
MACS2-IDR	4,946	9,337	35,033	3,991	5,584	23,274	6,364	6,078
SPP-IDR	4,861	12,160	40,006	5,095	5,042	25,470	7,074	6,794
diffReps	6,030	5,781	29,317	3,992	3,474	3,499	4,270	3,642
edgeR-basic	6,790	14,463	43,443	6,962	6,643	15,731	7,581	16,397
edgeR-plus	7,868	14,057	40,841	8,116	9,667	13,303	7,315	12,426

Table S4. Significance cut-offs for ChIP-Seq programs involved:

Program	Significance cutoff
PePr	p-value < 1e-5
MACS	p-value < 1e-5
ZINBA	Posterior Probability> 0.95
SICER	FDR < 1e-2
MACS2-IDR	*optimum set
SPP-IDR	*optimum set
edgeR	p-value < 1e-4
diffReps	p-value < 1e-4
DiffBind	FDR < 0.1

*See program parameters for details.

Table S5. Window sizes and shift sizes estimated for each dataset with PePr

Dataset	Window size (bp)	Shift sizes (bp)
NRSF	200	48, 46 (rep1,rep2)
ATF4	160	48, 40, 39 (rep1, rep2, rep3)
CTCF	260	86, 78 (rep1, rep2)
GABP	240	33, 47 (rep1, rep2)
NRF1	200	41, 54 (rep1, rep2)
SMC3	220	60, 53 (rep1, rep2)
USF1	200	40, 53 (rep1, rep2)
USF2	200	68, 55 (rep1, rep2)
H3K27me3	336	92, 91, 88, 90 (HPV(-) rep1,2, HPV(+) rep1,2)

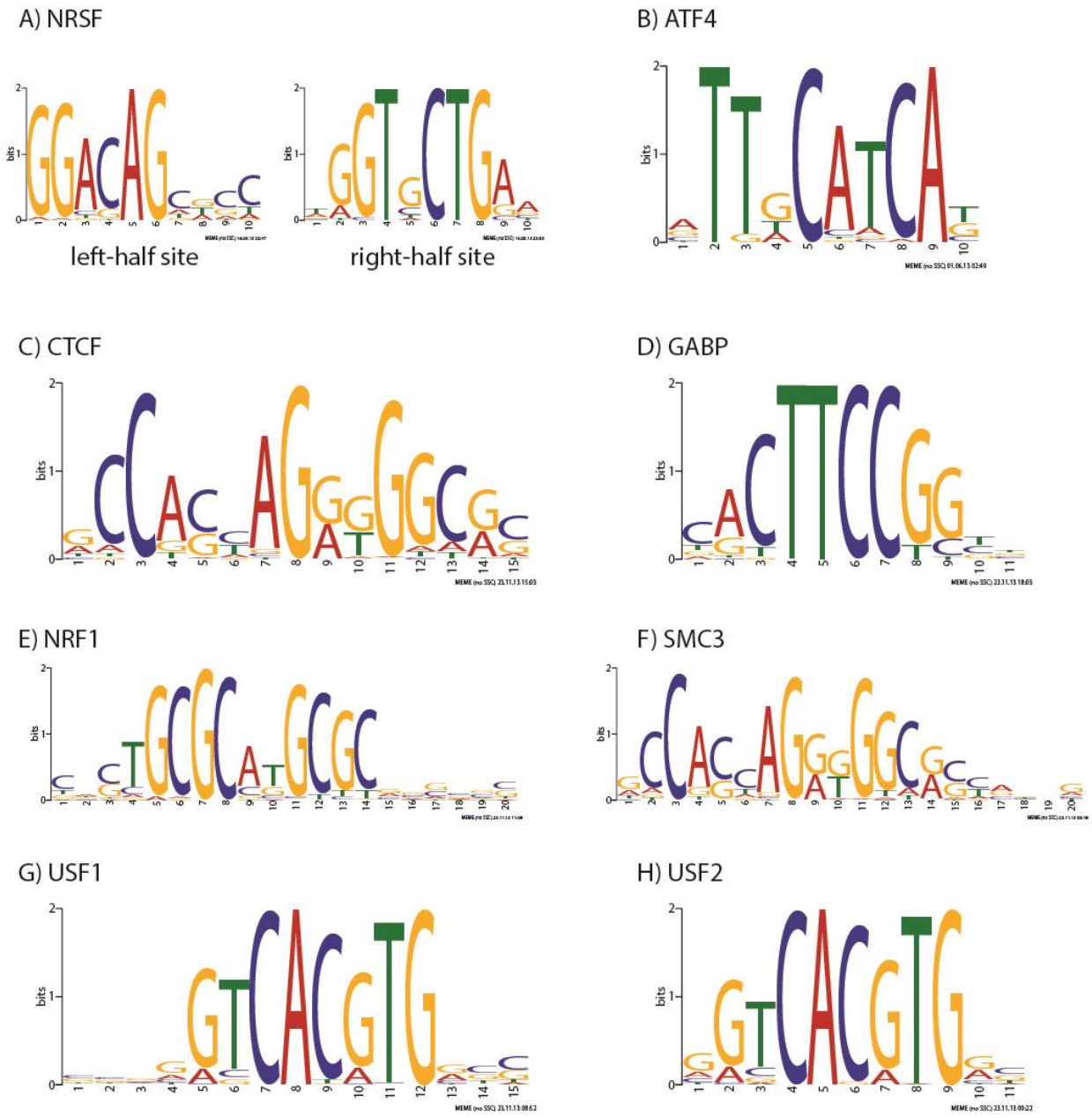


Figure S1. Motif logos for all TFs used in our comparisons. (A) Motif logo identified by MEME for NRSF data. NRSF binding sites have variable spacing between the two halves of the motif. (B-H) Motif logo identified by MEME for ATF4 (B), CTCF (C), GABP (D), NRF1 (E), SMC3 (F), USF1 (G) and USF2 (H).

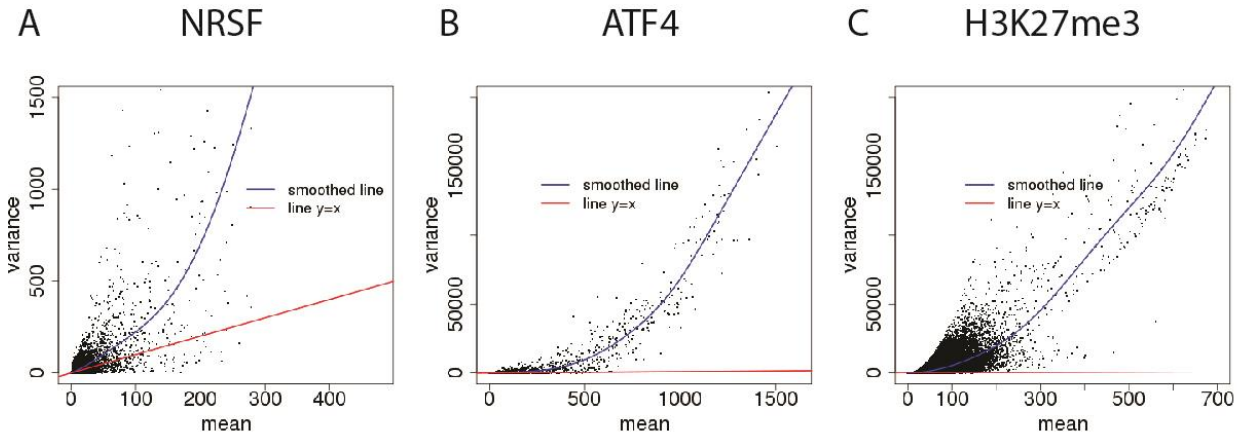


Figure S2. Extra-variance beyond that of the Poisson distribution is observed in ChIP-Seq data. Plot of mean versus variance estimates for windows across the genome in (A) NRSF ChIP-Seq data with two replicates (window size of 200bp), (B) ATF4 ChIP-Seq data with three biological replicates (window size of 160bp), and (C) H3K27me3 ChIP-Seq data from squamous cell carcinoma cell lines, with four replicates (window size of 340bp). The red line indicates the expected fit based on the Poisson distribution. The blue line is the fitted curve estimated using cubic smoothing spline.

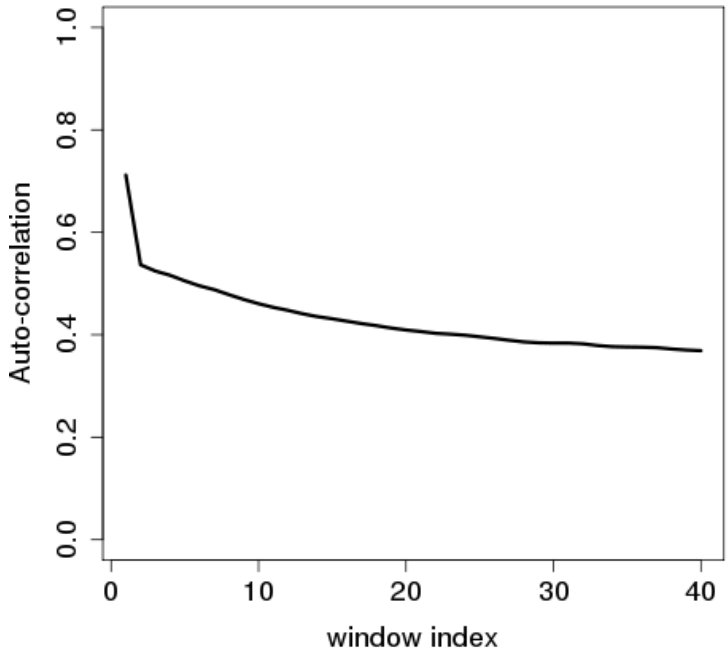


Figure S3. H3K27me3 data show a high autocorrelation of the dispersion parameters estimated for nearby windows. The genome was split into non-overlapping windows of 336 bp (Optimal window size estimated by PePr) and the dispersion parameter for each window was estimated. The autocorrelation of the dispersion parameters of the windows separated by (i-1) windows showed a correlation coefficient greater than 0.4 for to the range of 10 - 20 windows apart.

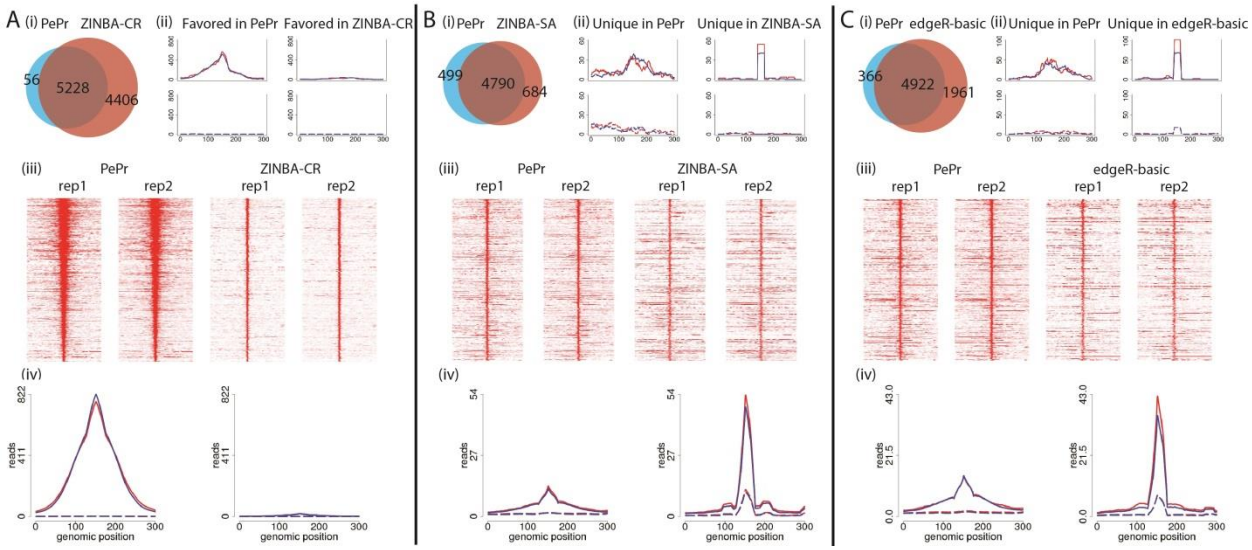


Figure S4. Comparison of PePr to ZINBA-CR (A) ZINBA-SA (B) and edgeR-basic (C) on NRSF data. (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group.

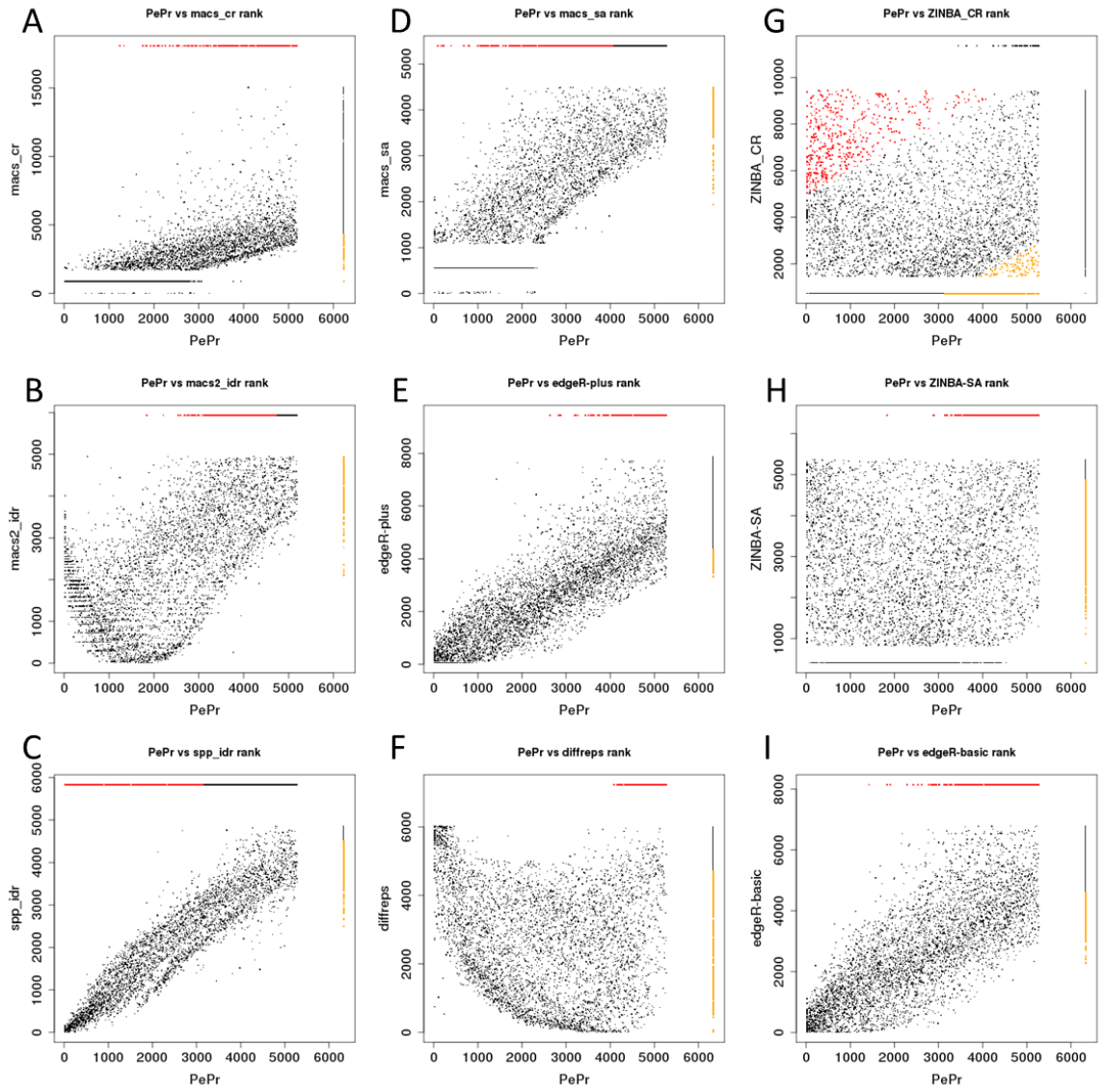


Figure S5. Rank comparisons between PePr and the alternative approaches on NRSF data. Rank comparisons between PePr and (A) MACS-CR (Pearson's r=0.73), (B) MACS2-IDR (r=0.65), (C) SPP-IDR (r=0.93), (D) MACS-SA (r=0.79), (E) edgeR-plus (r=0.84), (F) diffReps (r=-0.25), (G) ZINBA-CR (r=0.14), (H) ZINBA-SA (r=0.16), and (I) edgeR-basic (r=0.78). The peaks are ranked by the significance for each program. The points located at the top of each plot are PePr-unique peaks, and the points on the right margin of each plot are unique peaks for the alternative approach. Red and orange points refer to subsets of PePr-unique and alternative-approach-unique peaks which were used in the enrichment signal and motif occurrence comparisons.

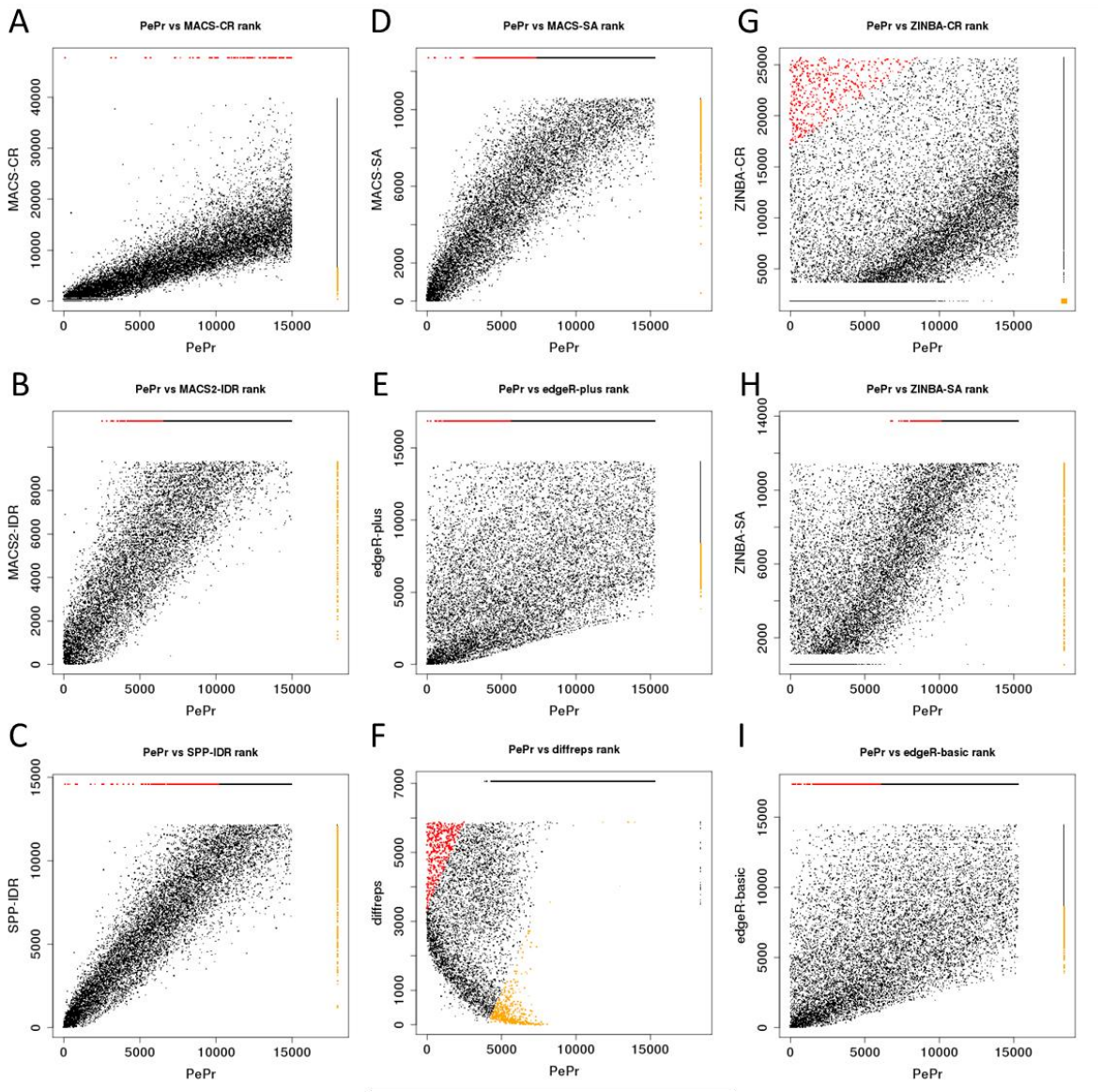
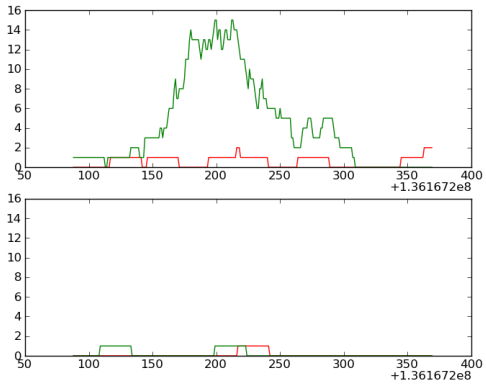
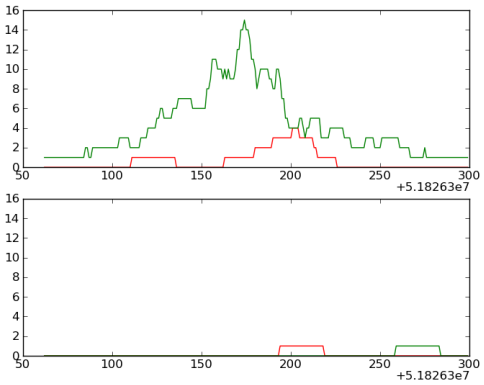


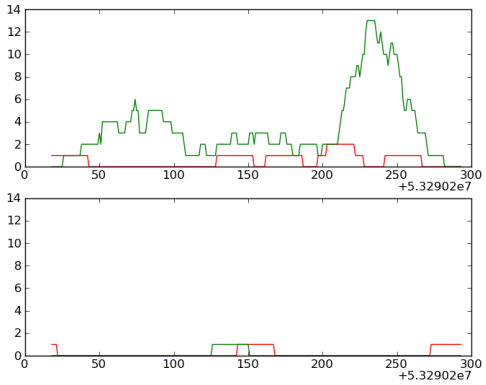
Figure S6. Rank comparisons between PePr and the alternative approaches on ATF4 data. Rank comparisons between PePr and (A) MACS-CR (Perason's r=0.82), (B) MACS2-IDR (r=0.82), (C) SPP-IDR (r=0.90), (D) MACS-SA (r=0.86), (E) edgeR-plus (r=0.51), (F) diffReps (r=-0.14), (G) ZINBA-CR (r=0.28), (H) ZINBA-SA (r=0.68), and (I) edgeR-basic (r=0.56). The peaks are ranked by the significance for each program. The points located at the top of each plot are PePr-unique peaks, and the points on the right margin of each plot are unique peaks for the alternative approach. Red and orange points refer to subsets of PePr-unique and alternative-approach-unique peaks which were used in the enrichment signal and motif occurrence comparisons.



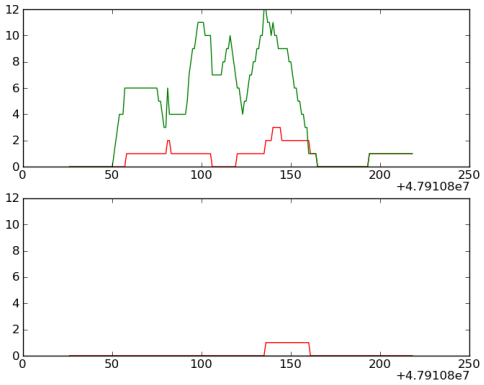
chr8:136,167,288-136,167,570



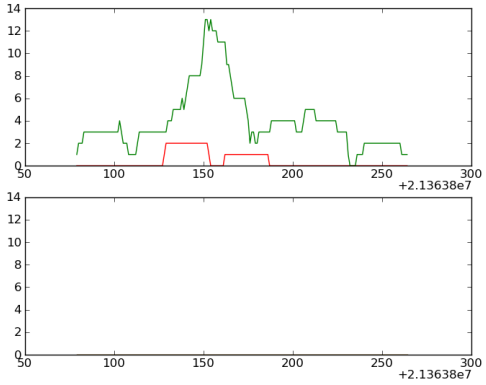
chr20: 51,826,362-51,826,600



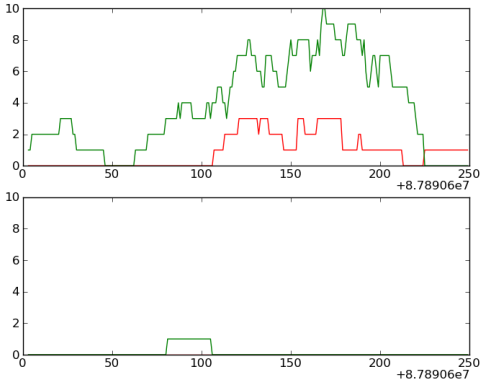
chr8: 53,290,218-53,290,494



chrX: 47,910,826-47,911,019

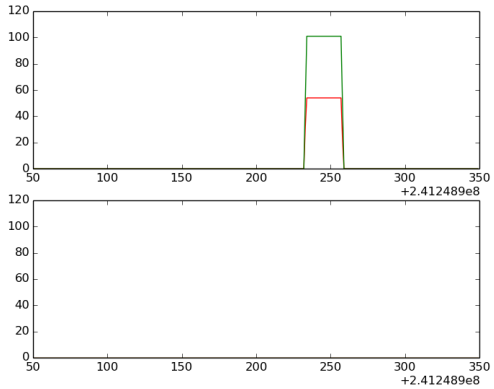


chr11: 21,363,879-21,364,065

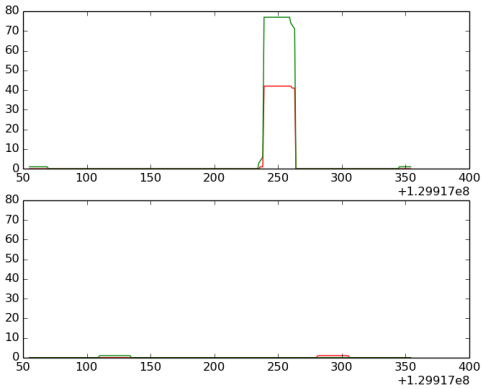


chr4: 87,890,603-87,890,850

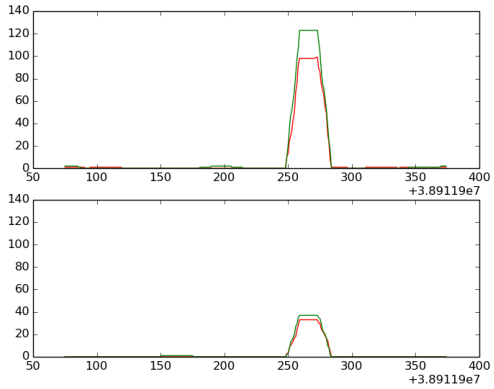
Figure S7. Additional examples of peak profiles uniquely identified by all CR- and IDR-based approaches (MACS, MACS2, ZINBA, and SPP) in NRSF data. For each plot, the x-axis denotes the genomic coordinates in base pairs, and the y-axis shows the read coverage. The top half of the plot shows the ChIP samples with green and red representing the two different replicates. The bottom half of the plot shows the input control samples in matching colors, with no line indicating zero reads across the region.



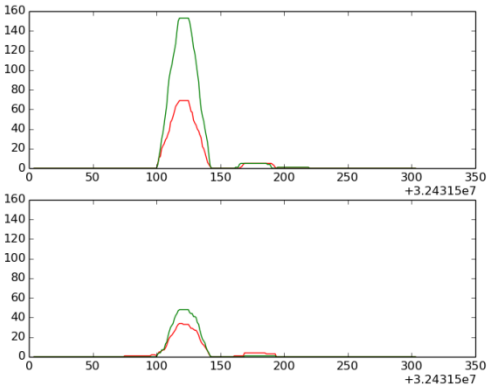
chr1: 241,248,950 – 241,249,250



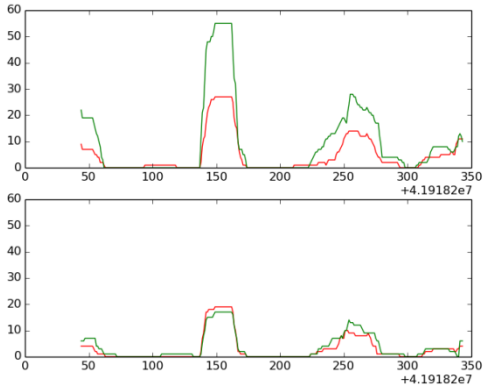
chr2: 129,917,055 – 129,917,355



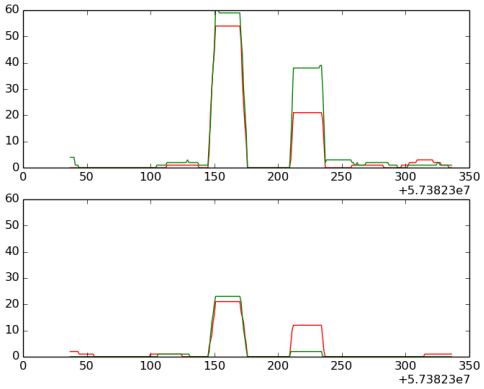
chr10: 38,911,975-38,912,275



chr19: 32,431,504 – 32,431,804



chr10: 41,918,244 – 41,918,544



chrY: 57,382,337 – 57,382,637

Figure S8. Examples of peaks uniquely identified by diffReps (not found by PePr) that are strongly spiked at the size of read length and/or exhibit the same peak profile for both ChIP-Seq and control samples. For each plot, the x-axis denotes the genomic coordinates in base pairs, and the y-axis shows the read coverage. The top half of the plot shows the ChIP samples with green and red lines representing the two different replicates, whereas the bottom half of the plot shows the input control samples in matching colors.

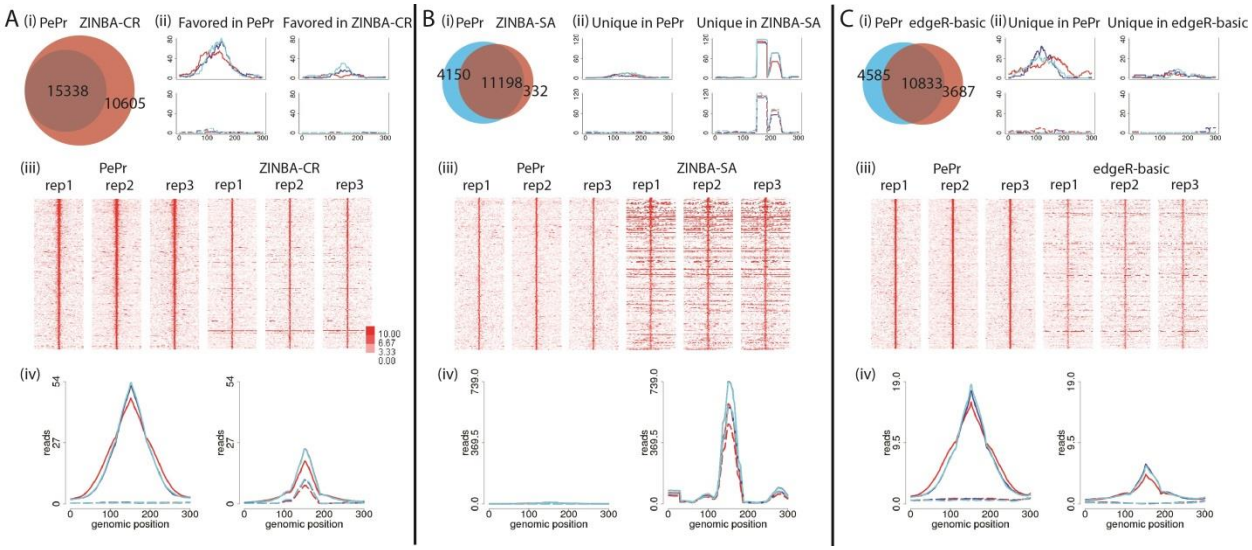


Figure S9. Comparison of PePr to ZINBA-CR (A), ZINBA-SA (B) and edgeR-basic (C) on ATF4 data. (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group.

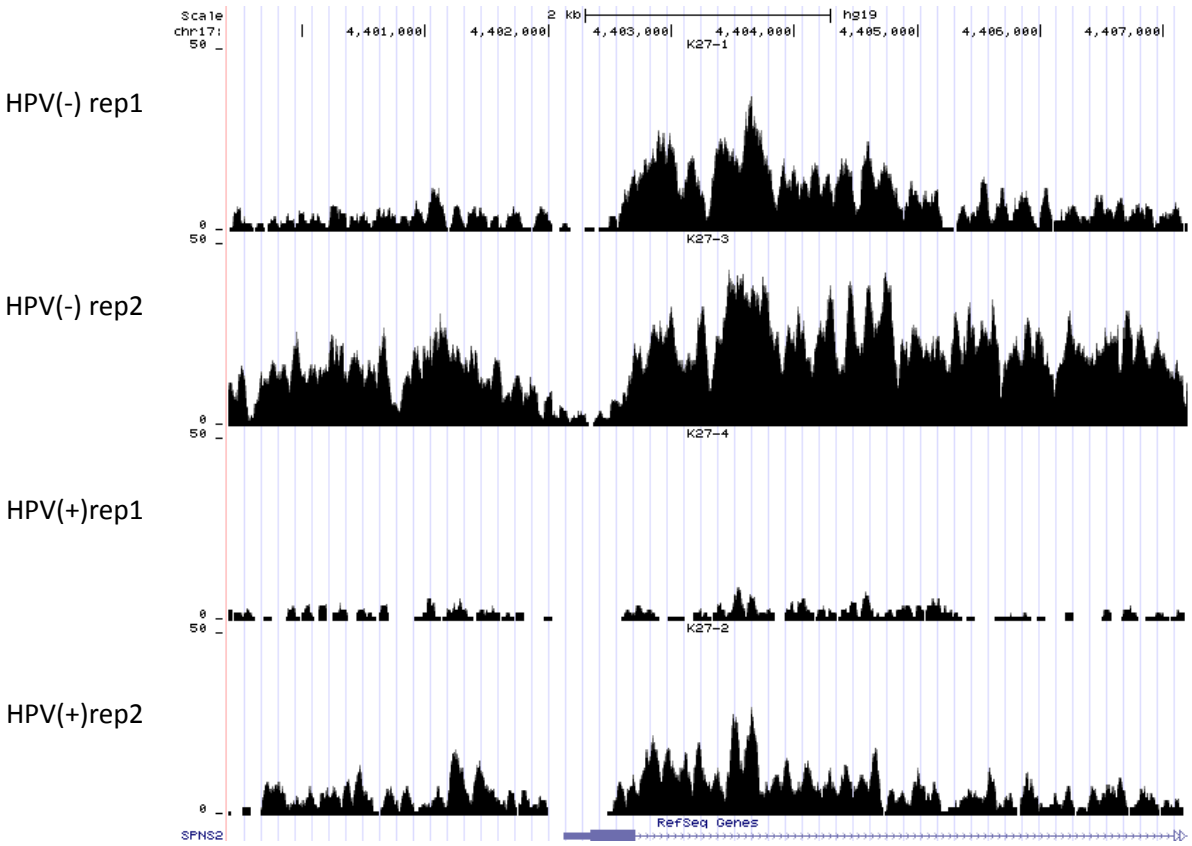


Figure S10. Example of an H3K27me3 enriched region showing high variation of ChIP-Seq signals across samples. Each profile represents one ChIP-Seq sample, with the x-axis and y-axis denoting chromosomal location and read coverage respectively.

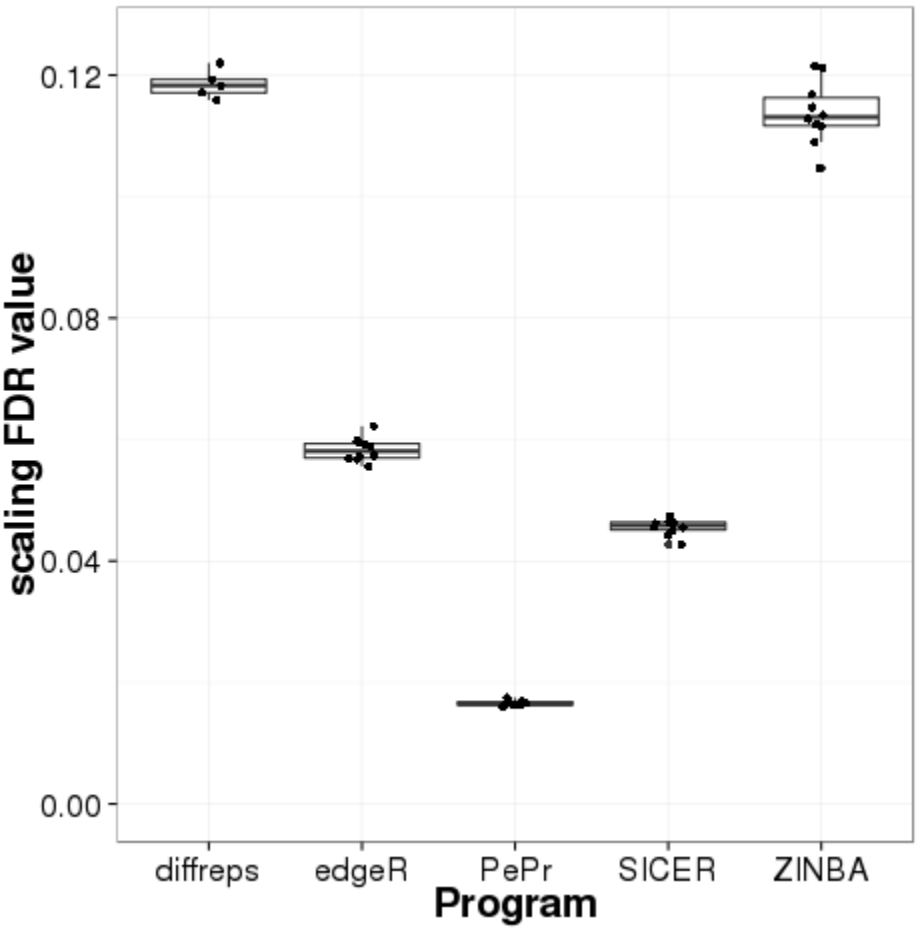


Figure S11. A scaling FDR analysis of the H3k27me3 dataset shows PePr was most robust to differences in read coverage level. The scaling FDR was calculated for PePr, ZINBA, SICER, diffReps, and edgeR on the H3K27me3 data as described in the main text. PePr had the lowest scaling FDR estimate of the methods tested.

References

Landt, S.G.*, et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, *Genome Res*, **22**, 1813-1831.

Liang, K. and Keles, S. (2012) Normalization of ChIP-seq data with control, *BMC Bioinformatics*, **13**, 199.

Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data, *Genome Biol*, **11**, R25.

Zhang, Y.*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS), *Genome Biol*, **9**, R137.